Heatmaps

In this exercise we are going to explore different datasets and determine the appropriate color scheme for each. In the workshop, the shiny widget showed that colour changes represent sharp perceptual breaks. Therefore, you should only use a color break, if it represents something. For instance, two different data populations, positive and negative values in the data.

INSERT INFORMATION ABOUT HOW TO CREATE HEATMAP - How to manipulate your data should be a lesson - needs to be in form Var 1 Var 2 Value

Lets take a look at the first dataset below.

`DatasetOne` contains average temperature data for different cities around the world.

```
DatasetOne <- read_excel("average-monthly-temperatures-acr.xls")
DatasetOne
```

```
## # A tibble: 36 x 25
##    Month               `Alice Springs ~ `Brisbane Airpo~ `Cairns Airport`
##    <dttm>                         <dbl>            <dbl>            <dbl>
##  1 2008-01-01 00:00:00             32.1             24.4             27.6
##  2 2008-02-01 00:00:00             28.4             23.8             27
##  3 2008-03-01 00:00:00             25.9             22               25.6
##  4 2008-04-01 00:00:00             20.7             19.4             23.9
##  5 2008-05-01 00:00:00             15.8             17.1             22.3
##  6 2008-06-01 00:00:00             12.8             16.4             22.3
##  7 2008-07-01 00:00:00             13               14.6             20.7
##  8 2008-08-01 00:00:00             12.7             14.4             21.3
##  9 2008-09-01 00:00:00             21.6             18.9             23.9
## 10 2008-10-01 00:00:00             25.5             20.2             25
## # ... with 26 more rows, and 21 more variables: `Sydney Airport` <dbl>,
## #   `Wagga Airport` <dbl>, Feuerkogel <dbl>, `Salzburg-Flughafen` <dbl>,
## #   Uccle <dbl>, `Belize / Phillip Goldston Intl. Airport` <dbl>, `San
## #   Joaquin` <dbl>, `Inukjuak, Que` <dbl>, Beijing <dbl>, Guangzhou <dbl>,
## #   `Bordeaux / Merignac` <dbl>, `Marseille / Marignane` <dbl>, `Budapest
## #   / Lorinc` <dbl>, Reykjavik <dbl>, `Dublin Airport` <dbl>, Tokyo <dbl>,
## #   `Luxembourg / Luxembourg` <dbl>, `Christchurch (1864-2011)` <dbl>,
## #   `Zurich Town / Ville.` <dbl>, `Huron Regional Airport` <dbl>, `Colonia
## #   (1951-2011)` <dbl>
```

Lets take a look at the distribution of the data as well as the maximum and minimum temperatures to see if there are any perceptual breaks in the data.

There are a few ways to explore the dataset to see if you can notice any perceptual breaks in the data.

1. Look at the summary of the whole dataset.

```
summary(DatasetOne)
```

```
##      Month                       Alice Springs Aerodrome
##  Min.   :2008-01-01 00:00:00   Min.   :11.00
##  1st Qu.:2008-09-23 12:00:00   1st Qu.:15.22
##  Median :2009-06-16 00:00:00   Median :21.90
##  Mean   :2009-06-16 10:40:00   Mean   :21.32
##  3rd Qu.:2010-03-08 18:00:00   3rd Qu.:27.65
##  Max.   :2010-12-01 00:00:00   Max.   :32.10
##
##  Brisbane Airport M. O Cairns Airport  Sydney Airport  Wagga Airport
##  Min.   :14.40         Min.   :20.70   Min.   :12.20   Min.   : 7.30
##  1st Qu.:17.18         1st Qu.:23.00   1st Qu.:15.40   1st Qu.:10.72
##  Median :20.50         Median :25.25   Median :18.55   Median :15.50
##  Mean   :20.28         Mean   :24.87   Mean   :18.41   Mean   :16.12
##  3rd Qu.:23.65         3rd Qu.:26.93   3rd Qu.:22.02   3rd Qu.:21.15
##  Max.   :25.50         Max.   :28.00   Max.   :24.10   Max.   :26.50
##
##    Feuerkogel     Salzburg-Flughafen     Uccle
##  Min.   :-6.700   Min.   :-3.300     Min.   :-0.70
##  1st Qu.:-2.525   1st Qu.: 3.725     1st Qu.: 6.25
##  Median : 4.050   Median : 9.100     Median :10.55
##  Mean   : 3.694   Mean   : 9.361     Mean   :10.43
##  3rd Qu.: 8.975   3rd Qu.:15.700     3rd Qu.:16.18
##  Max.   :13.500   Max.   :20.600     Max.   :19.40
##
##  Belize / Phillip Goldston Intl. Airport  San Joaquin    Inukjuak, Que
##  Min.   :22.20                            Min.   :22.90   Min.   :-26.400
##  1st Qu.:25.10                            1st Qu.:26.23   1st Qu.:-15.950
##  Median :27.50                            Median :27.15   Median : -3.200
##  Mean   :26.81                            Mean   :26.70   Mean   : -4.258
##  3rd Qu.:28.60                            3rd Qu.:27.70   3rd Qu.:  5.625
##  Max.   :29.60                            Max.   :29.00   Max.   : 14.700
##  NA's   :1                                NA's   :2
##     Beijing        Guangzhou     Bordeaux / Merignac Marseille / Marignane
##  Min.   :-4.80   Min.   :11.60   Min.   : 3.50       Min.   : 4.80
##  1st Qu.: 1.90   1st Qu.:18.12   1st Qu.: 8.95       1st Qu.: 9.60
##  Median :14.95   Median :23.60   Median :13.25       Median :15.00
##  Mean   :13.06   Mean   :22.66   Mean   :13.26       Mean   :15.35
##  3rd Qu.:23.02   3rd Qu.:27.93   3rd Qu.:18.70       3rd Qu.:21.30
##  Max.   :28.60   Max.   :29.90   Max.   :22.10       Max.   :26.10
##
##  Budapest / Lorinc   Reykjavik      Dublin Airport       Tokyo
##  Min.   :-2.000    Min.   :-0.200   Min.   : 0.000   Min.   : 5.500
##  1st Qu.: 5.625    1st Qu.: 1.050   1st Qu.: 5.650   1st Qu.: 9.875
##  Median :12.050    Median : 4.450   Median : 8.300   Median :17.100
##  Mean   :11.492    Mean   : 5.586   Mean   : 8.891   Mean   :16.667
##  3rd Qu.:19.050    3rd Qu.:10.125   3rd Qu.:12.700   3rd Qu.:23.150
##  Max.   :23.600    Max.   :13.000   Max.   :16.000   Max.   :29.600
##                                     NA's   :1
##  Luxembourg / Luxembourg Christchurch (1864-2011) Zurich Town / Ville.
##  Min.   :-2.500          Min.   : 5.200           Min.   :-1.900
##  1st Qu.: 4.475          1st Qu.: 7.775           1st Qu.: 4.125
```

```
##  Median : 9.200          Median :11.500          Median : 9.800
##  Mean   : 9.403          Mean   :11.592          Mean   : 9.433
##  3rd Qu.:16.025          3rd Qu.:15.175          3rd Qu.:15.675
##  Max.   :20.500          Max.   :18.500          Max.   :20.000
##
##  Huron Regional Airport Colonia (1951-2011)
##  Min.   :-11.400        Min.   :10.10
##  1st Qu.: -1.975        1st Qu.:13.50
##  Median :  7.800        Median :16.80
##  Mean   :  7.033        Mean   :17.61
##  3rd Qu.: 17.900        3rd Qu.:22.60
##  Max.   : 24.100        Max.   :25.70
##                         NA's   :3
```

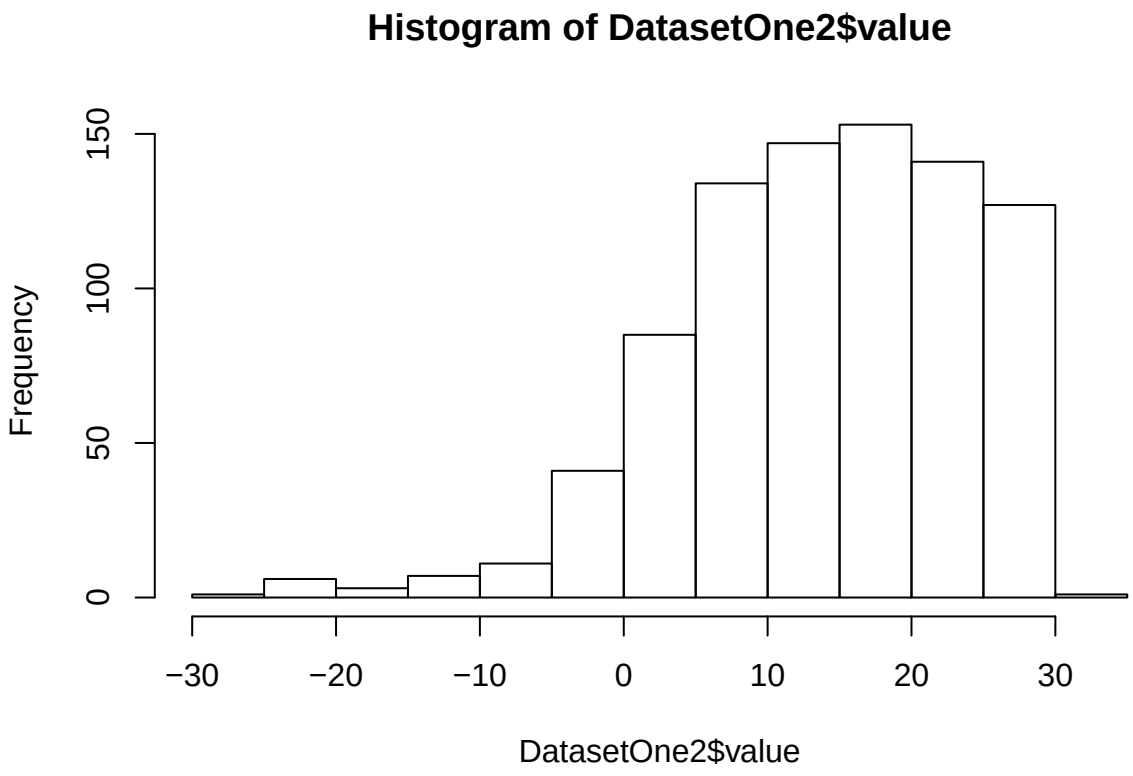
2. Use visualisation to explore the distribution of the dataset.

In order to visualise the distribution of the data and to generate a heatmap, we need to manipulate the data into the form . To do this we need to use `dplyr` packages function gather. shown below. The maniplated dataset is stored as `DatasetOne2`.

```
DatasetOne2 <- gather(DatasetOne, City, value, c(2:ncol(DatasetOne)))
```

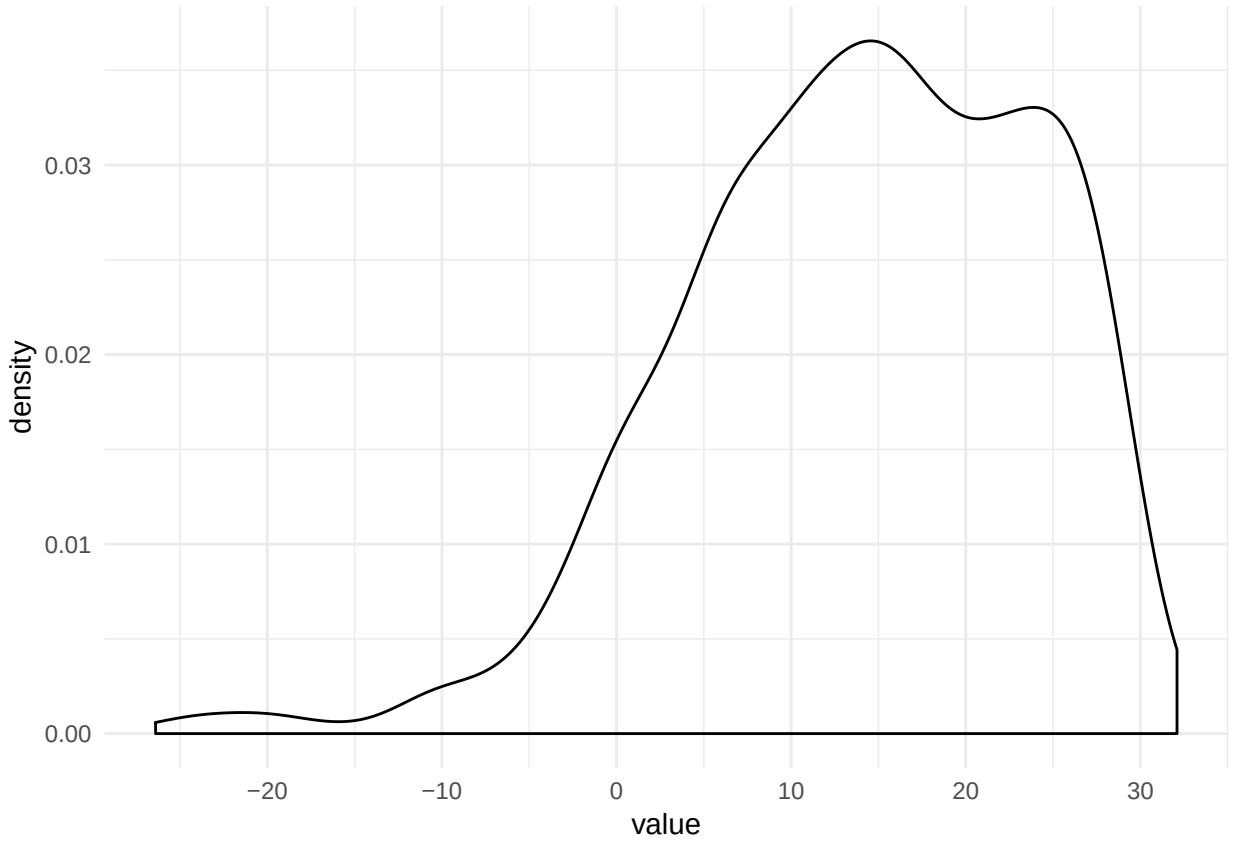
- *Note: There are two different ways to do this. A histogram or Density plot. Use your preferred method here.*

```
hist(DatasetOne2$value)
```




```
ggplot(DatasetOne2, aes(x = value)) + geom_density()
```

```
## Warning: Removed 7 rows containing non-finite values (stat_density).
```



Using the information that you have just found, determine an appropriate color scheme for this heatmap. There are three ways to change the color sheme of the heatmap.

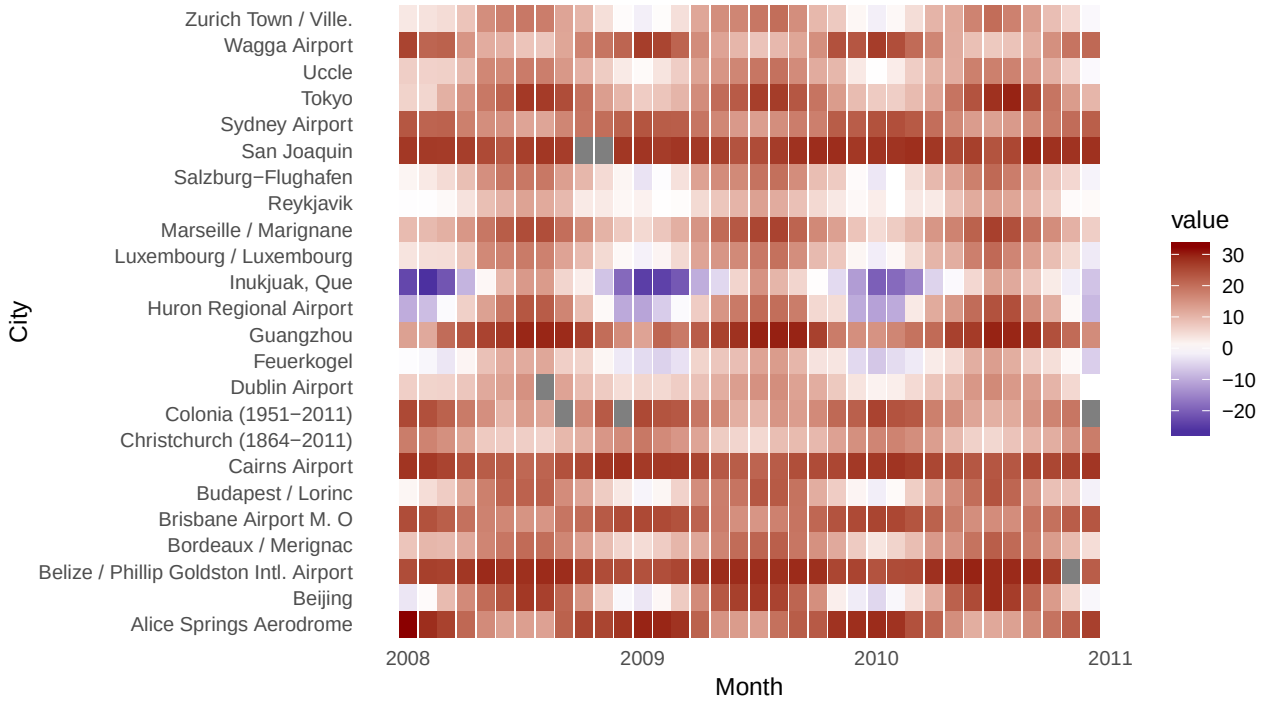
- scale_fill_gradient() - Two-color Gradient
- scale_fill_gradient2() - Gradient with a middle color and two colors that diverge from it
- scale_fill_gradientn() - Gradient with n colors, equally spaced.

The heatmap command has been completed below, run it to take a look at the base heatmap. Then update the color scheme using one of the functions above.

Have a go at choosing your color scheme in the answer sheet notebook before scrolling down to take a look at the answer on the next page.

Answer: Taking a look at the summary and distribution of the data we can notice on perceptual break in the data, negative and positive temperatures. With this in mind, `scale_fill_gradient2()` was used to generate a colour scheme with two different colors and a midpoint at 0.

```
DatasetOne2 <- gather(DatasetOne, City, value, c(2:ncol(DatasetOne)))
ggplot(data = DatasetOne2, aes(x = Month, y = City)) +
      geom_tile(aes(fill = value)) +
      theme(
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background = element_blank()
      ) + scale_fill_gradient2(low = "darkblue", midpoint = 0, mid = "white", high = "darkred")
```



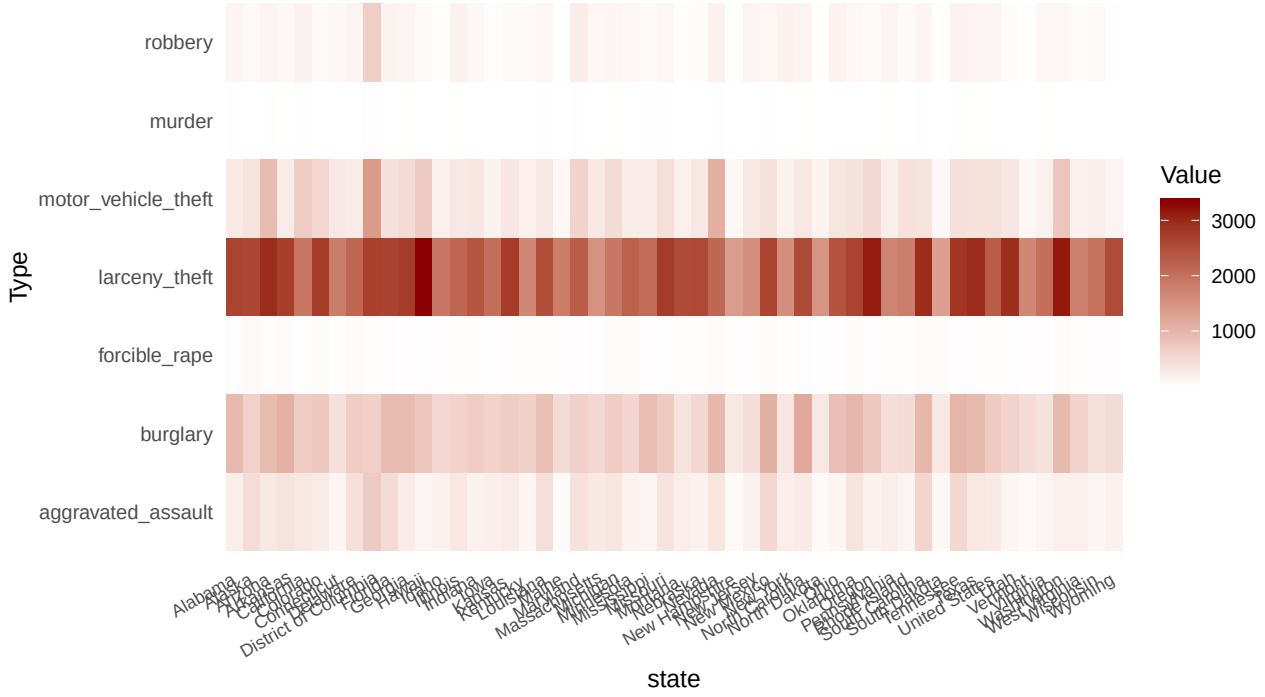
Three more datasets have been provided in the answer sheet notebook for you to take and look at and determine the correct color scheme to use. The proposed answers to these datasets are shown on the next page. Be sure to take a go for yourself before you take a look and follow the process to explore your data before creating your heatmap.

`DatasetTwo` - contains level the highest level of education obtains for Canadians over the age of 15 by province. The data was obtained from "https://open.canada.ca/data/en/dataset/4e0ddb90-e4ad-421b-b074-ad6fb6a96dae."
`DatasetThree` - contains the crime statistics in America by state and type of crime.
`DatasetFour` - contains responses from a student survey taken from "https://computerstats.wordpress.com/2016/11/26/correlation-heatmaps-r-and-excel/". We will use this data to create a correlation matrix.

The proposed color schemes are on the next page. Give the exercises a go before taking a look.

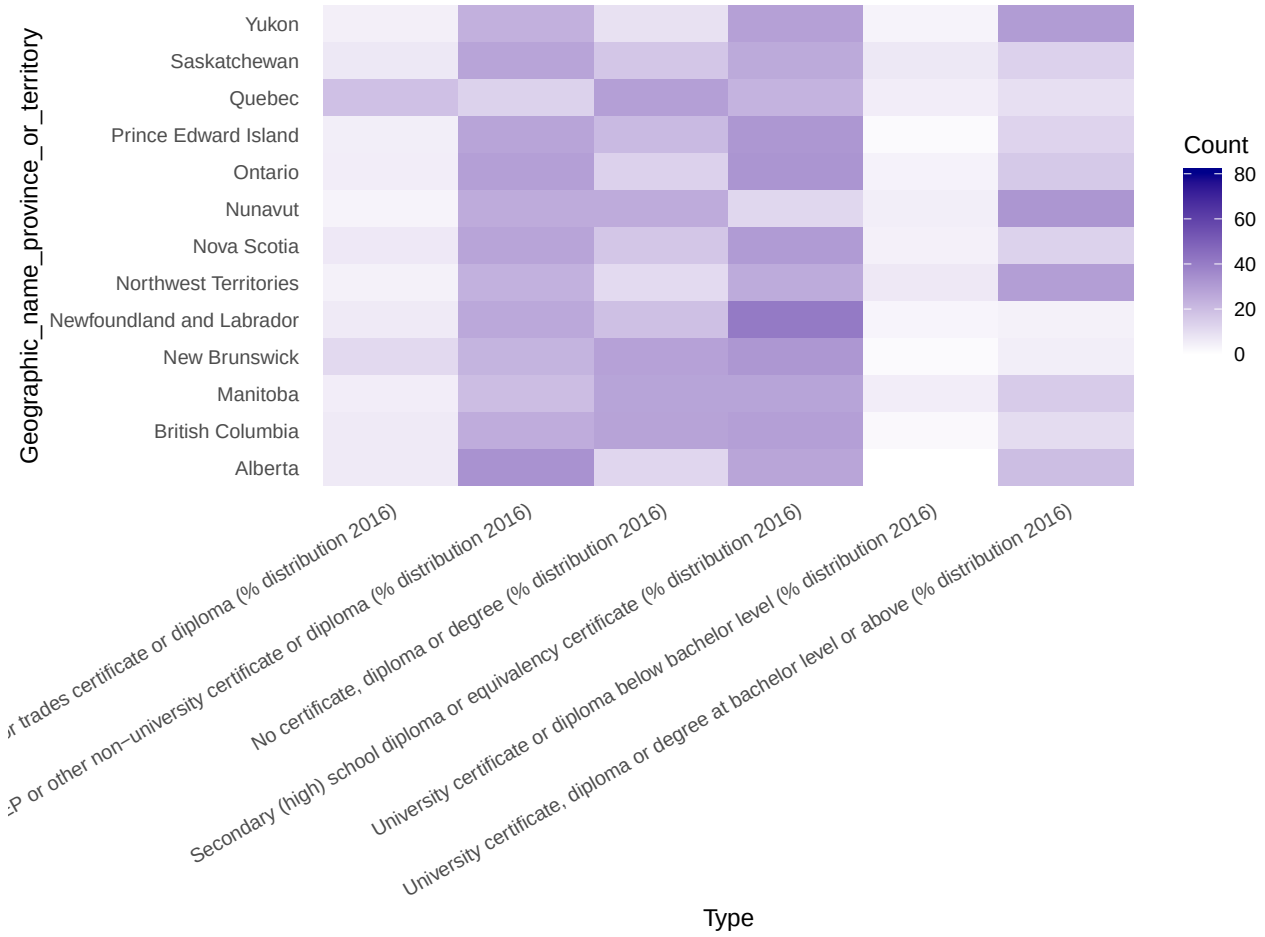
```
DatasetThree <- read.csv("crimeRatesByState-formatted.csv", sep = ",", header = TRUE)
DatasetThree2 <- gather(DatasetThree, Type, Value, 2:ncol(DatasetThree))
```

```
ggplot(data = DatasetThree2, aes(x = state, y = Type)) +
      geom_tile(aes(fill = Value)) +
      theme(
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background = element_blank(),
        axis.text.x = element_text(angle = 30, hjust = 1, size = 8)
        ) + scale_fill_gradient2(low = "white", high = "darkred")
```



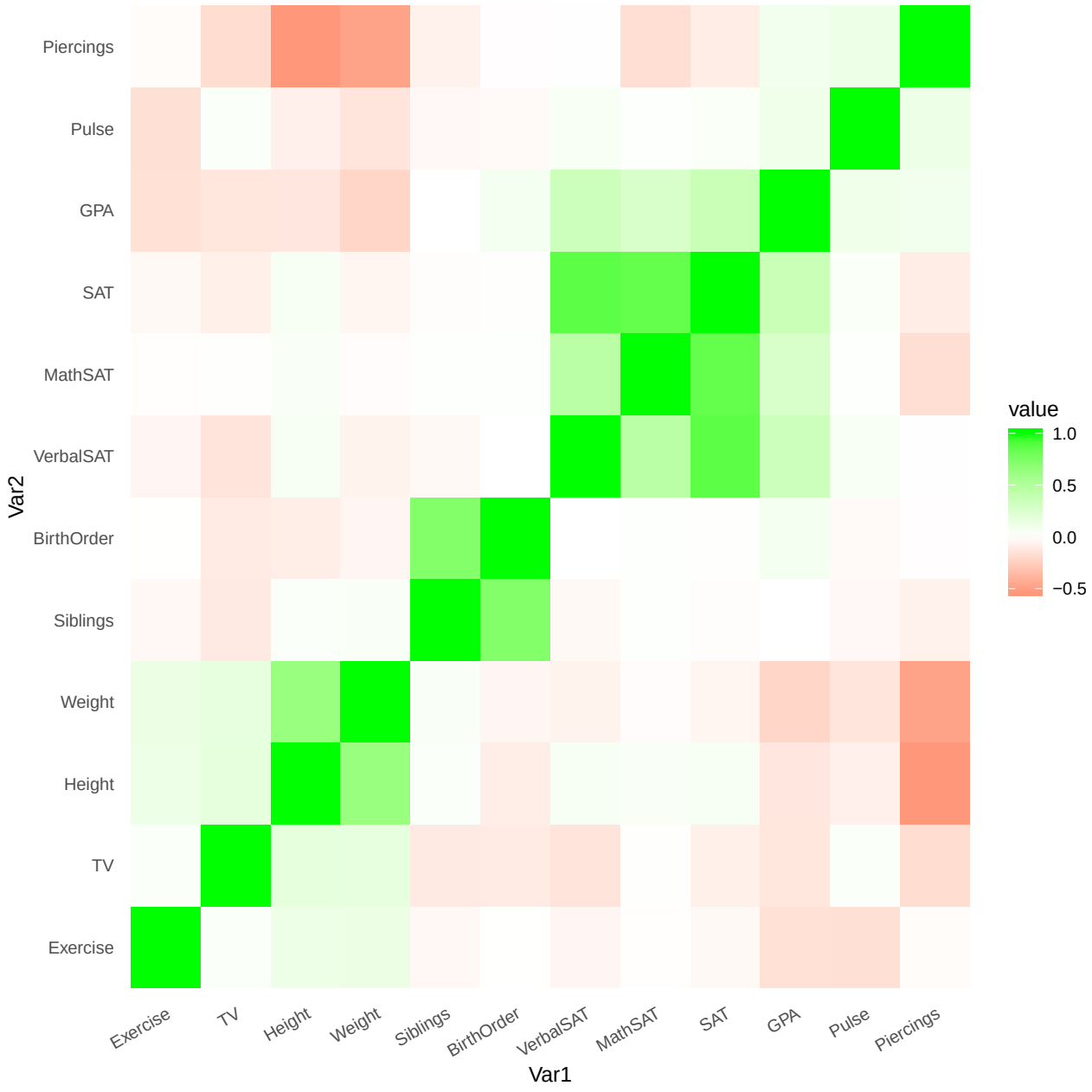

```
DatasetTwo <- read_excel("canadaeducation2.xls")
DatasetTwo <- gather(DatasetTwo, Type, Count, 2:ncol(DatasetTwo))
```

```
ggplot(data = DatasetTwo, aes(x = Type, y = Geographic_name_province_or_territory)) +
      geom_tile(aes(fill = Count)) +
      theme(
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background = element_blank(),
        axis.text.x = element_text(angle = 30, hjust = 1)
        ) + scale_fill_gradient2(low = "white", high = "darkblue")
```




```
original <- read.csv("StudentSurvey.csv", header = TRUE)
original1 <- original[,6:ncol(original)]
DatasetFour <- cor(original1, use="na.or.complete")
DatasetFour <- melt(DatasetFour)
```

```
ggplot(data = DatasetFour, aes(x = Var1, y = Var2)) +
      geom_tile(aes(fill = value)) +
      theme(
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background = element_blank(),
        axis.text.x = element_text(angle = 30, hjust = 1)
        ) + scale_fill_gradient2(low = "red", high = "green")
```


Piercings
Pulse
GPA
SAT
MathSAT
VerbalSAT
BirthOrder
Siblings
Weight
Height
TV
Exercise
Var2
Exercise
TV
Height
Weight
Siblings
BirthOrder
VerbalSAT
MathSAT
SAT
GPA
Pulse
Piercings
Var1
value
1.0
0.5
0.0
−0.5